

Methodology

Open Access

EST-PAC a web package for EST annotation and protein sequence prediction

Yvan Strahm*¹, David Powell*¹ and Christophe Lefèvre^{†1,2}

Address: ¹Victorian Bioinformatics Consortium, Monash University, Clayton Vic 3800, Australia and ²Department of Zoology, the University of Melbourne, Melbourne Vic 3010, Australia

Email: Yvan Strahm* - yvan.strahm@gmail.com; David Powell* - powell@csse.monash.edu.au; Christophe Lefevre - c.lefevre@zoology.unimelb.edu.au

* Corresponding authors †Equal contributors

Published: 12 October 2006

Received: 23 May 2006

Source Code for Biology and Medicine 2006, 1:2 doi:10.1186/1751-0473-1-2

Accepted: 12 October 2006

This article is available from: <http://www.scfbm.org/content/1/1/2>

© 2006 Strahm et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

With the decreasing cost of DNA sequencing technology and the vast diversity of biological resources, researchers increasingly face the basic challenge of annotating a larger number of expressed sequences tags (EST) from a variety of species. This typically consists of a series of repetitive tasks, which should be automated and easy to use. The results of these annotation tasks need to be stored and organized in a consistent way. All these operations should be self-installing, platform independent, easy to customize and amenable to using distributed bioinformatics resources available on the Internet.

In order to address these issues, we present EST-PAC a web oriented multi-platform software package for expressed sequences tag (EST) annotation. EST-PAC provides a solution for the administration of EST and protein sequence annotations accessible through a web interface. Three aspects of EST annotation are automated: 1) searching local or remote biological databases for sequence similarities using Blast services, 2) predicting protein coding sequence from EST data and, 3) annotating predicted protein sequences with functional domain predictions. In practice, EST-PAC integrates the BLASTALL suite, EST-Scan2 and HMMER in a relational database system accessible through a simple web interface. EST-PAC also takes advantage of the relational database to allow consistent storage, powerful queries of results and, management of the annotation process. The system allows users to customize annotation strategies and provides an open-source data-management environment for research and education in bioinformatics.

Findings

An expressed sequences tag (EST) is the result of sequencing a portion of a cDNA clone derived from an mRNA [1]. EST sequencing is especially useful for gene discovery in species lacking a draft genome sequence. Used in conjunction with genomic sequencing, ESTs have been used to characterize gene expression products, define intron-exon boundaries, find genes or gene locations [2,3], and ana-

lyze splice variation and polymorphism [4]. In many instances EST annotation allows the identification or functional prediction of cloned inserts. These clones can be used subsequently in the laboratory for the experimental characterization of bioactivity. The annotation of EST libraries requires a number of repetitive tasks (see [5] for a review) that are easily automated: database searches, translations into peptides and, functional annotation of

translation products with probabilistic model searches such as the protein family database (Pfam) [6]. Efforts in the open source and in the academic community have been made to provide the scientific community with on line services, examples of which are PipeOnline, [7], EST-PAGE [8], or complete packages such as ESTannotator [9], ESTAP [10], PartiGene [11], and Prot4EST [12]. However, these packages often have restrictive system dependencies, do not always allow extensive data mining and, may not always be available for download and customization. Furthermore, few packages allow real sequence management where users can decide to build, store and use complex filters for sequence similarity searches with criteria based on previous results. This facility offers more flexibility for the annotation process, updating and the optimal use of often limited computational resources. Here, we propose a flexible interface and a PHP Hypertext Preprocessor programming language framework for annotation data mining.

The management and updating of sequence annotations is facilitated using EST-PAC and the source code is accessible for customization of the web interface.

The core of EST-PAC consists of an open source relational database management system that uses Structured Query Language (MySQL) and a number of PHP programs. EST-PAC uses open source software; MySQL 4 or 5 [13] for database management and, PHP 4 or PHP 5 [14] for web development. PHP allows the storage and management of ESTs using web pages. User login is available for visualization and query only or, with additional privileges to run annotation tools. Sequences in FASTA format are loaded into the database through a web interface and annotation tasks can be requested. A set of continuously running programs checks the database and extracts sequence to be processed using the BLASTALL suite [15], ESTScan2 [16,17] or, HMMER [18] against the Pfam database [19].

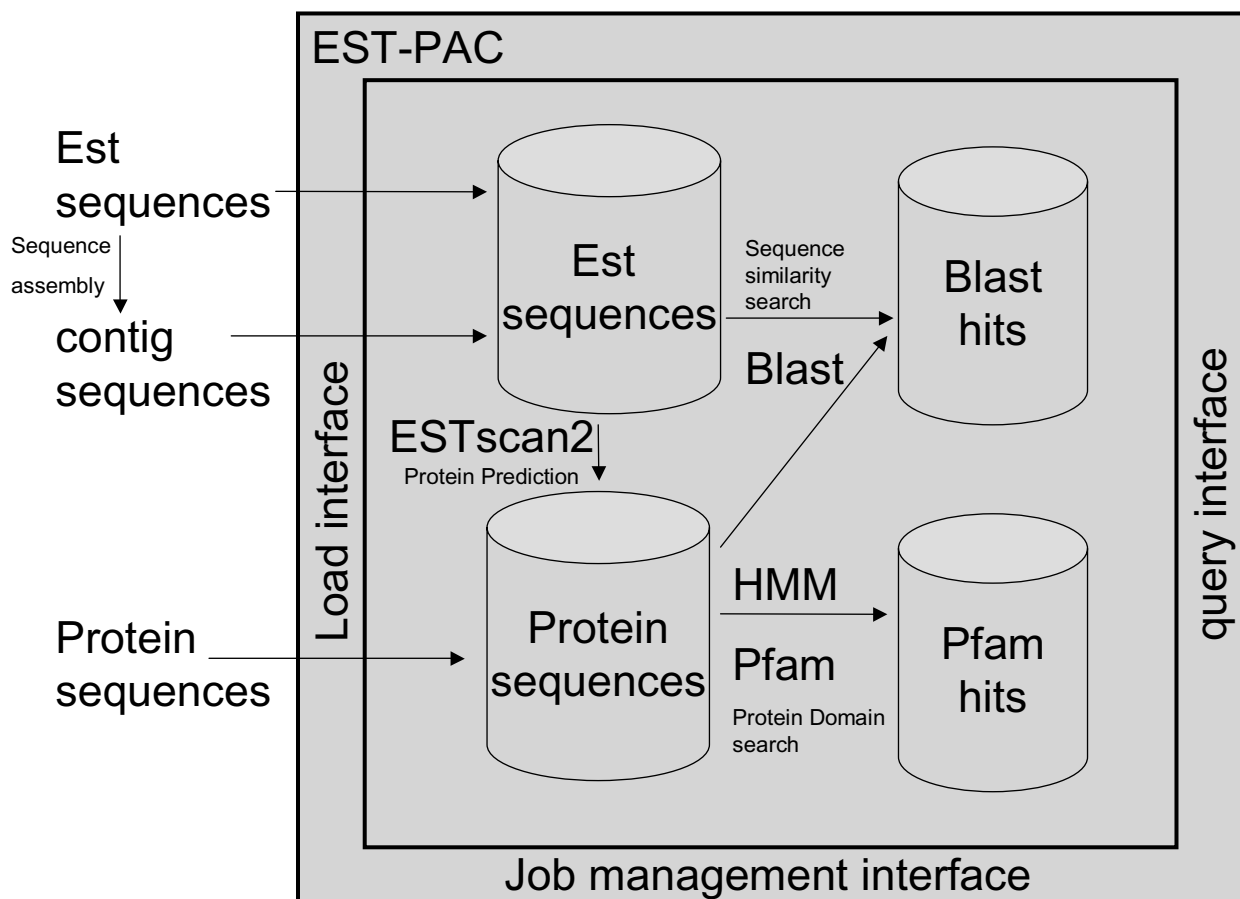


Figure 1

Workflow and interfaces available in EST-PAC. EST-PAC provides a web interface for the management, storage and querying of sequences and results from annotation tools such as BLASTALL, EST-Scan2 and HMMER.

The coding content of the EST can be evaluated with the Hidden Markov Model approach of ESTScan2 and the predicted translation products can then be compared against protein sequence databases. A report can be obtained from a web query page. As all results are stored in a relational database, users are able to query on every value returned by the annotation process. An interface is also available to assist the construction and storage of database queries. In addition to the public databases which can be downloaded and installed locally or accessed through web based blast services such as NCBI, users have the possibility to create their own databases from EST-PAC in order to make more precise and relevant comparisons. We have tried to restrict the programming in EST-PAC to the PHP language. However, Perl 5.8.1 was used to integrate ESTScan2. The usage of Perl is limited to this and a full installation of Perl or BioPerl [20] is not necessary. System specific configuration of EST-PAC has been kept to a minimum. However, some indispensable set-up is needed. First, MySQL should be running and MySQL administrator login and password are required during installation. Second the configuration file (`config.inc.php`) should be edited to reflect the environment where the package is installed. The user must indicate where in the operating system the package and auxiliary programs are located as well as the name of the database to be created. It is also possible to create multiple databases by specifying database names, usernames, and passwords. These parameters will be passed automatically at the time of database creation. The BLASTALL programs can be used either locally or remotely. With the remote option, blast jobs will be sent to the blast server at NCBI [21] in compliance with NCBI remote access policy. Access to other resources can also be implemented using the EST-PAC framework.

EST-PAC is distributed in a comprehensive PHP package for the integration of all the third party bioinformatics annotation pipeline program components. Installation of the auxiliary components (NCBI tool suite version 2.2.10, HMMER version 2.3.2 ESTSCAN version 2.0b, MySQL version 5, and PHP) is also necessary. EST-PAC scripts can be downloaded [22]. Installation and configuration instructions are available here for MacOSX, Windows or Linux systems. In addition, precompiled version of ESTScan2 for MacOSX, Linux Fedora Core 3, Centos 4 and Windows are also available at this URL.

EST-PAC is a software package for the annotation of EST data particularly geared towards laboratories with limited bioinformatics resources and expertise. It provides a basic package to install a database system for the management of a complete suite of third party annotation software components, integrated into a simple and powerful web interface. Mainly built in PHP, the source code is easily

accessible to developers for customization and evolution. EST-PAC does not currently address directly the assembly of EST sequences. It is however possible to assemble EST data independently and straightforward to annotate the contig sequences obtained after the assembly process. Additional support for EST assembly data storage and visualization is in development. Furthermore, EST-PAC already provides a workbench to cluster ESTs onto reference sequence data sets when available, for example, from public genome annotation data. Finally, usage of EST-PAC is not restricted to EST sequences and any type of nucleotide or protein sequences can be loaded for the management of sequence analysis results. This allows the compilation, storage and management of a diversity of customized sequence databases for the analysis of EST or other sequence libraries. In conclusion, EST-PAC provides an open framework for rapid prototyping of data mining and on-line visualization of sequence data, presenting an expandable data-management environment for research and education in bioinformatics.

Availability

Project name: EST-PAC.

Project home page: <http://vbc.med.monash.edu.au/~yvan/est-pac>.

Operating system(s): multiplatform; Programming language: PHP 4 or 5;

Other requirements: mySQL, BLASTALL suite, ESTScan2 and HMMER.

A web page with a fully functional instance of EST-PAC has been set up for demonstration purpose. Software download and installation instructions are also available under GPL.

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

YS was involved in the conceptualisation, development and testing of the package. DP was involved in the programming of the core functionality of EST-PAC. CL supervised the work, implemented the graphic components and, prepared the manuscript.

Acknowledgements

The authors would like to thank Torsten Seemann for expert advise, Marc Liyanage for providing the MacOSX Apache/PHP/MySQL package <http://www.entropy.ch>, the Apache friends for the Windows XAMP package <http://www.apachefriends.org>, and Claudio Lottaz and Christian Iseli for allowing us to incorporate ESTScan2 in this package.

References

1. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF: **Complementary DNA sequencing: expressed sequence tags and human genome project.** *Science* 1991, **252**:1651-1656.
2. Dias Neto E, Correa RG, Verjovski-Almeida S, Briones MR, Nagai MA, da Silva W, Zago MA, Bordin S, Costa FF, Goldman GH, Carvalho AF, Matsukuma A, Baia GS, Simpson DH, Brunstein A, de Oliveira PS, Bucher P, Jongeneel CV, O'Hare MJ, Soares F, Brentani RR, Reis LF, de Souza SJ, Simpson AJ: **Shotgun sequencing of the human transcriptome with ORF expressed sequence tags.** *Proc Natl Acad Sci USA* 2000, **97**:3491-3496.
3. Hillier LD, Lennon G, Becker M, Bonaldo MF, Chiapelli B, Chissoe S, Dietrich N, DuBuque T, Favello A, Gish W, Hawkins M, Hultman M, Kucaba T, Lacy M, Le M, Le N, Mardis E, Moore B, Morris M, Parsons J, Prange C, Rifkin L, Rohlfing T, Schellenberg K, Marra M: **Generation and analysis of 280,000 human expressed sequence tags.** *Genome Res* 1996, **6**:807-828.
4. Wistow G, Bernstein SL, Wyatt MK, Fariss RN, Behal A, Touchman JW, Bouffard G, Smith D, Peterson K: **Expressed sequence tag analysis of human RPE/choroid for the NEIBank Project: over 6000 non-redundant transcripts, novel genes and splice variants.** *Mol Vis* 2002, **8**:205-220.
5. Shivashankar HN, Gasser RB, Ranganathan S: **A hitchhiker's guide to expressed sequence tags (ESTs) and their bioinformatics analysis.** *Briefings Bioinf* in press.
6. Bateman A, Birney E, Cerruti L, Durbin R, Ewlinger L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2002, **30**:276-280.
7. Ayoubi P, Jin X, Leite S, Liu X, Martajaja J, Abduraham A, Wan Q, Yan W, Misawa E, Prade RA: **PipeOnline 2.0: automated EST processing and functional data sorting.** *Nucleic Acids Res* 2002, **30**:4761-4769.
8. Matukumalli LK, Grefenstette JJ, Sonstegard TS, Van Tassell CP: **EST-PAGE. Managing and analyzing EST data.** *Bioinformatics* 2004, **20**:286-288.
9. Hotz-Wagenblatt A, Hankeln T, Ernst P, Glatting KH, Schmidt ER, Suhai S: **ESTAnnotator: A tool for high throughput EST annotation.** *Nucleic Acids Res* 2003, **31**:3716-3719.
10. Mao C, Cushman JC, May GD, Weller JW: **ESTAP. An automated system for the analysis of EST data.** *Bioinformatics* 2003, **19**:1720-1722.
11. Parkinson J, Anthony A, Wasmuth J, Schmid R, Hedley A, Blaxter M: **PartiGene. Constructing partial genomes.** *Bioinformatics* 2004, **20**:1398-1404.
12. Wasmuth JD, Blaxter ML: **prot4EST: translating expressed sequence tags from neglected genomes.** *BMC Bioinformatics* 2004, **5**:187.
13. **MySQL** [<http://www.mysql.com>]
14. **PHP** [<http://www.php.net>]
15. McGinnis S, Madden TL: **BLAST: at the core of a powerful and diverse set of sequence analysis tools.** *Nucleic Acids Res* 2004, **32**:W20-25.
16. Iseli C, Jongeneel CV, Bucher P: **ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences.** *Proc Int Conf Intell Syst Mol Biol* 1999:138-148.
17. Lottaz C, Iseli C, Jongeneel CV, Bucher P: **Modeling sequencing errors by combining Hidden Markov models.** *Bioinformatics* 2003, **19**(Suppl 2):103-112.
18. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**:755-763.
19. **Pfam** [<http://www.sanger.ac.uk/Software/Pfam/>]
20. **BioPerl** [<http://bio.perl.org>]
21. **NCBI BLAST** [<http://www.ncbi.nlm.nih.gov/blast/>]
22. **EST-PAC** [<http://vbc.med.monash.edu.au/~yvan/est-pac>]

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

